

# **Workflow and Diagnostics**

**Dean N. Williams and Katherine Evans**

**On behalf of the ACME Workflow Team**

**Accelerated Climate Modeling For Energy  
Proposal Review**

# ACME is a grand challenge of complex coupled model development

## Science Needs

## Solution

**Wide range of model runs and workflow types across the project**

Define use cases: 1) Model Development (DEV); (2) Exploratory Model Runs (EXP); (3) Production Model Runs (PROD)

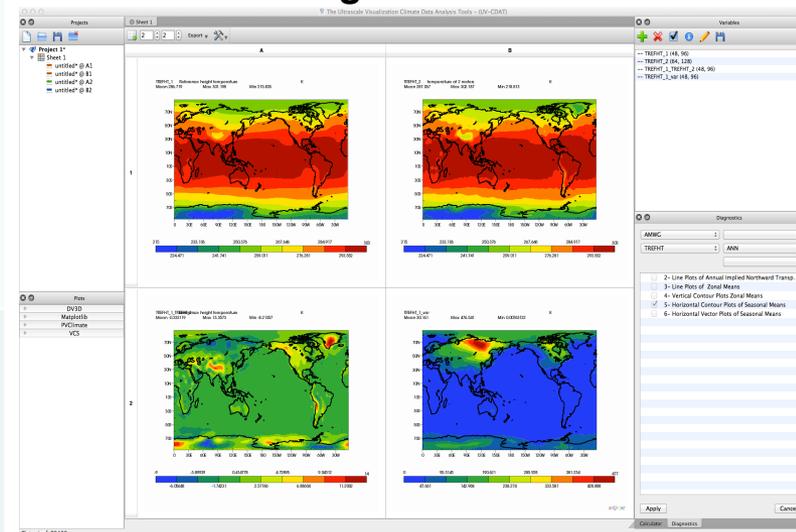
**Capture and record suites of runs and their settings during model development**

Automated provenance and archiving

**Quickly evaluate coupled model behavior**

Diagnostics of the coupled system within one software system

## Diagnostics



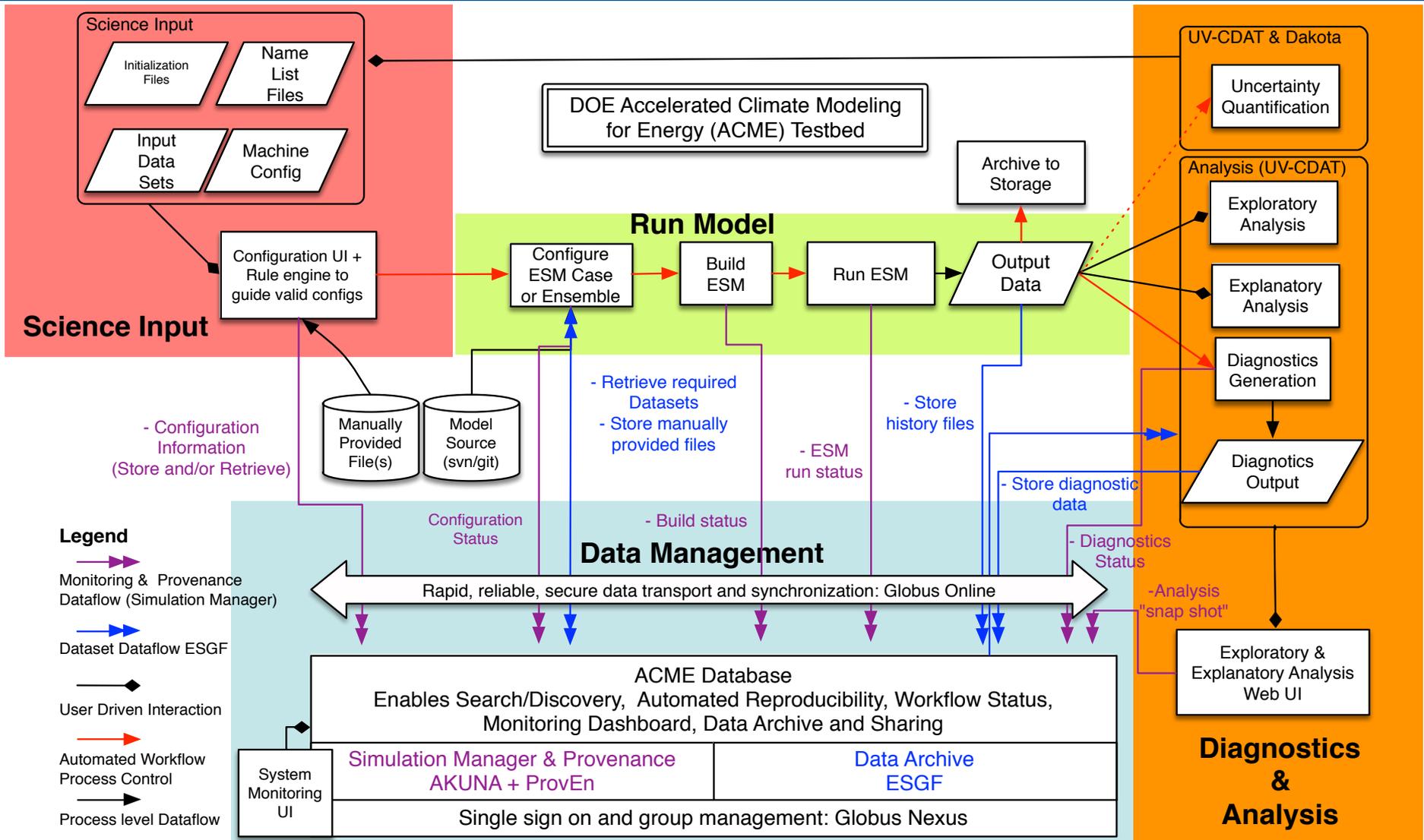
# Major challenges for the ACME end-to-end system

<b>Challenges</b>	<b>Description</b>
<b>Installation</b>	Software must be adapted to multiple hardware platforms and operating systems located throughout the ACME
<b>Heterogeneous Data Sets</b>	The same infrastructure must also allow scientists to access and compare data sets from multiple sources, including from observational satellite and instrument sources
<b>Analysis, Diagnostics, and Visualizations</b>	The generation of new and improved analysis, diagnostics, and visualization techniques for the better model development and evaluation
<b>Server-side and In Situ Computing</b>	Server-side and in situ computation is necessary as the increase in data size and complexity of algorithms lead to data-intensive, compute-intensive challenges for ACME diagnostics, UQ, analysis, model metrics, and visualization

# Testing and execution framework tailored for ACME

Features	Description	Impact
<b>Infrastructure</b>	Creates a flexible, extensible infrastructure for future ACME efforts and related DOE projects, automates laborious, repetitive simulation data tasks to improve productivity	Heightens productivity and user experience
<b>Data Sharing</b>	Supports broad data sharing within ACME project teams and with scientific collaborations; including NGEE, ARM, CDIAC, etc.	Accelerates model development and result dissemination
<b>Provenance</b>	Enables reproducibility, archiving and reuse of high-volume simulation data, provenance captures set up, execution and analysis details coupled with standard metadata creation, annotation, and forums for group discussions and sharing of any part of a workflow	Increases reproducibility, productivity and credibility of collaboration
<b>Model set up and execution</b>	Rule based support for model setup, specialized collaborative portal with a checklist for approvers of new model setups before job launching, Links to User interface and infrastructure for job submission	Enables new users to be effective quickly, allows control over model set ups in distributed, collaborative teams
<b>User Interface</b>	Specialized software when needed to enable web job submission, running, monitoring, and debugging capabilities on several HPC centers	One stop shop to all needed capabilities, increases productivity, reproducibility
<b>Ensemble and Automated Runs</b>	Job launching interface for submission of hundreds of production runs and enabling specialized monitoring of multiple ensemble, automated runs	Increases productivity

# End-to-end workflow and data infrastructure architecture



# Accelerating the science

Deliverable	Description
<b>Faster Data Movement</b>	Leverage the use of Globus transfer and sharing, integrated with ESGF; UV-CDAT at the LCF facilities and laboratories; work with ESnet to achieve good network performance throughput for rapid and secure data transport
<b>Improved Diagnostics</b>	Incorporate standard diagnostics in UV-CDAT for all sub-model components, including more advanced diagnostics, UQ algorithms, ensembles, etc.
<b>Exploratory and Analysis Visualization</b>	Incorporate additional exploratory and analysis (i.e., EDEN) capabilities based on climate scientist and model developer requirements
<b>Improved Interface</b>	Refine fundamental workflow issues with user interface in order to increase scientific productivity due to hard-to-use software
<b>Faster Analysis and Diagnostics (Parallelism)</b>	Incorporate parallel capabilities into diagnostics framework for climatology generation and batch processing; add parallel support in UV-CDAT analysis and visualization frameworks where required
<b>Provenance Capture</b>	Using Akuna and ProvEn, capture provenance of the full workflow allowing reproducibility of model configuration, execution, and diagnostics
<b>Tracking and Feedback</b>	Display the current status of each workflow run
<b>Automated Job Launching</b>	Complete the specialized collaborative test bed portal with a checklist for approvers of new model setups before workflow job launching
<b>Improve Usability</b>	User support for infrastructure, including online training material, on-site training courses, and ongoing user support

# Workflow Team

Team	Personnel
<b>T3 Task Lead</b>	Rachana Ananthakrishnan – ANL Eric Blau – ANL Charles Doutriaux – LLNL Katherine Evans – ORNL John Harney – ORNL Kerstin Kleese-Van Dam – PNNL Carina Lansing – PNNL Elo Leung – LLNL Po-Lun Ma – PNNL Ben Mayer – ORNL Renata McCoy – LLNL Jeffrey Painter – LLNL Galen Shipman – ORNL Brian Smith – ORNL Chad Steed – ORNL
<b>T4 Task Lead</b>	D. Wang – ORNL Dean Williams – LLNL Jin-Ho Yoon – PNNL

Workflow Config interface showing a sequence of steps: Code, Config, Inputs, Build, Run, Output. A sidebar on the left lists various models and groups like Atmosphere Model, Ocean Model, and Ice Model. The main area is titled 'Config'.

ACME Workflows UV-CDAT ESGF Data Transfer Globus AKUNA CoG Log In Register

### Workflow Home

user info

AKUNA ESGF

username: mattben openid: [input]

password: [input] password: [input]

save cancel forget

### Available Templates

F1850\_g37 clone

### User Workflows

Privacy & Legal Notice • Site Issues

UV-CDAT+ Home Login Geo Diagnostics Tree

## Exploratory Analysis - Diagnostics Tree Viewer

Create Tree Bookmarks Saved Images

Tree Name: Bookmark2

Dataset: Variable: Time:

Submit

Tree structure showing variables like NPP, PBOT, PCO2, FSAT, FSH, and TLAQ with associated time periods (MAR, APR, MAY, JUNE, JULY).

UV-CDAT Live

Project1

Sheet1

234.567790912

246.03940328

3.2158075684

Plot Types

UV-CDAT-EA Home Login Geo Diagnostics Tree Heatmap

Heatmap

Dataset: tropics\_warming\_th\_co2

Column Variables: Select options

Row Variables: Select options

Apply

Grid of heatmaps for variables: NPP, PBOT, PCO2, FSAT, FSH, Q2M.

localhost:8000/web-apps/climate-diagnostics/index.html

Jan 0151 [16.5, 90]

Plots: DV3D, Mapped to POC/Climate, VCS

Summary

Submit a Job

Package: AMWG

Obs: NCEP

Set: 5- Horizontal Contour Plots of Seasonal M...

Season: ANN

Variable: TREFHT

Auxiliary: [input]

Submit

Results

13171

TREFHT\_2\_ANN.nc

TREFHT\_2\_ANN.nc

TREFHT\_2\_ANN.nc

globus Manage Transfers Groups Support williams13

Transfer Files

Endpoint enter endpoint name Go

Path Go

Please select an endpoint above.

Please select an endpoint above.

more options Label This Transfer

This will be displayed in your transfer activity.

The Ultraviolet Visualization Climate Data Analysis Tools - UV-CDAT

Variables

AMWG 1

TREFHT 1 ANN 5

2-Line Plots of Annual Implied Northward Transp...

3-Line Plots of Zonal Means

4-Vertical Contour Plots, Zonal Means

5-Horizontal Contour Plots of Seasonal Means

6-Horizontal Vector Plots of Seasonal Means

Four panels showing global maps and time-series plots of temperature anomalies.

Panel 1: Tropics warming (0151-0161)

Panel 2: Tropics warming (0151-0161)

Panel 3: Tropics warming (0151-0161)

Panel 4: Tropics warming (0151-0161)

Time-series plot: 11.823922965003418

Plots: DV3D, Mapped to POC/Climate, VCS

# ACME Reviewers' Workflow and Diagnostics Questions

# Testing and execution framework tailored for ACME

## Questions

- Q1** The section on big data is spot on, but I wonder if the team has considered adopting some of the innovations in big data technologies coming from Amazon, Google, etc. Dollars for cyberinfrastructures/data management technologies are scarce in DOE/science, so leveraging investments that are designed to scale ‘infinitely’ seems worth considering, allowing ACME resources to focus on more domain-related problems.
- Q2** The output from ACME v1 simulations will apparently be accessible through the existing infrastructure of the Earth System Grid (ESG) Federation. This mechanism for data distribution has proven to be successful with past projects (e.g., delivery of CMIP3 and CMIP5 model output). However, as the proposers correctly point out, a "big data" problem is emerging with the increase in resolution and the number of ensembles being planned. There are real questions to be answered regarding the extensibility of ESG in the "Big Data" era. The current paradigm of "search, order, and download" datasets won't survive with the volumes of data that are expected. In the near future the applications and computing will have to be close to the data for inter-comparisons between products. ESG appears to be underfunded and incapable of such a transformation.
- Q3** Development of modern workflow and diagnostics needed for proper address formidable validation and uncertainty quantification challenges in an extreme scale computing environment is tremendously difficult. Since it is not realistic to “do everything” – especially for a 3-year activity with constrained funding -- what is the prioritization strategy that the ACME proposal team is planning to adopt?
- Q4** The proposal contains rather limited discussion of how the model and its modules will be tested against reality. I see this as a critical part of the science that could flow from this work. Ideally, the ability to simulate observable quantities is a requirement. For example, satellite observations of full atmosphere CO2 concentration is in the works. These could provide valuable feedback on the CO2 fluxes and atmospheric circulation simulated in the model, but this requires that the model include the capacity to simulate CO2 concentration. There are many other possible observation/simulation pairs that could be used for testing. Treatment of this area in the proposal is somewhat vague. I’d like to hear some more specific thoughts on this topic.

# Testing and execution framework tailored for ACME

## Questions

- Q5** P92 discusses building upon a bunch of open source DOE efforts to compose an infrastructure for ACME and depicts this in Fig 6.1. I be interested to hear the team's thoughts on the complexity of (1) putting this infrastructure together – will changes need to be made to individual components or is it just 'plug and play' (2) have they assessed the robustness of each of the constituent tools (i.e., do they work as advertised – a risk?) (3) how the resulting system will be deployed/managed/evolved. Complex integration can lead to frail system architectures – there's trade-off here to consider perhaps?
- Q6** P93 Last paragraph discussed building flexibility into the tools to support different workflows, standalone operations, etc. Flexibility usually is desirable, but comes at the cost of more development and complexity, so I'd be interested to hear if the team has any specific use cases that would make it easier for us to understand the 'dimensions' of flexibility they have in mind? Very little comes for free in the software world ;)
- Q7** P94 last paragraph talks about provenance/metadata being managed by 4 tools!! Sounds complex, and makes me think there's possibly some compromise in this design, and more than likely complexity. Is it really needed, or are you just trying to keep everyone happy? What are the trade-offs?
- Q8** P96 talks generally about managing data resources. I wonder if the team has any concrete estimates of the volumes of data that will need to be managed in the 3 years, and then into the future? It's not clear to me if the solutions here are appropriate without some context on their usage/demands?
- Q9** P96 also talks about reusable Python modules for scripting, which is fine, and also providing REST based interfaces for 'user interfaces'. But I can call REST from the command line using curl – is there a chance to rationalize development effort here and just provide a single universal interface to these modules? It reduces effort, code volume and maintenance 'mortgages'.

# Testing and execution framework tailored for ACME

## Questions

- Q10** P97 briefly discusses service monitoring. In my experience this is crucial and often non-trivial, becoming a serious sink for unexpected efforts. It'd be interesting to hear the team's plans in this area.
- Q11** Section 6.1.4.4 discusses and IDE and Akuna. Are these two tools, or one with two functions. It's not clear from the description if you intend to develop an IDE?
- Q12** p92/93 perhaps expand on automation processes for testing.
- Q13** Workflow: the end to end workflow is a laudable goal, but is extraordinarily complex, as can be seen quite simply by scanning Fig 6-1-1. Has enough thought been given to fault tolerance within this system? What happens if some components in this system are down or offline? If there is a discussion somewhere, I missed it. An account of the current state of affairs in this workflow development would be useful.
- Q14** P97 6.1.4.2: Are there any plans to integrate NCL or PyNGL with UV-CDAT? Or will the existing NCL CESM diagnostics be fully ported to UV-CDAT? Is this a reasonable effort in view of the fact that the NCL diagnostics will continue to be developed by the CESM community?